View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Samara University

О возможностях пакета адаптивного регрессионного моделирования СПОР

Г.Р. Кадырова[a], Т.Е. Родионова[a]

[a] *Ульяновский государственный технический университет, 432027, ул. Северный Венец, 32, Ульяновск, Россия*

Аннотация

Представлен оригинальный статистический пакет «Система поиска оптимальных регрессий» (СПОР), позволяющий осуществлять высокоточное статистическое (регрессионное) моделирование процессов или явлений с последующим использованием моделей для прогноза выходных характеристик. Эффективность данной методологии, под которой понимается сокращение размерности модели и повышение точности определения ее параметров и прогноза, прямо пропорциональна размерности, степени зашумленности и мультиколлинеарности исходных данных, что позволяет считать ее применение для высокоточных расчетов перспективным математическим подходом.

Ключевые слова: регрессионное моделирование; прогнозирование; методы структурной идентификации; критерии качества модели; программный пакет

1. Введение

Программный пакет СПОР (система поиска оптимальных регрессий) является специализированной системой, реализующей стратегию адаптивного регрессионного моделирования (АРМ) [1].

На начальном этапе АРМ–подход предусматривает применение линейного регрессионного анализа (РА), предполагающий постулирование модели, оценивание параметров методом наименьших квадратов (МНК), статистический анализ модели и ее составляющих. В условиях соблюдения определенных предположений МНК–оценка $\hat{\beta}$ и прогноз $\hat{Y}$, считаются наилучшими линейными оценками (НЛО). К сожалению, упомянутые предположения в подавляющем случае нарушаются, что, приводя к искажению МНК-оценок $\hat{\beta}$, влечет за собой неконтролируемое увеличение случайных и систематических ошибок прогноза $\hat{Y}$.

На последующих этапах АРМ–подход предусматривает: - проверку соблюдения гипотез РА – МНК, ранжирование нарушений по степени искажения свойств наилучших линейных оценок или в зависимости от назначения модели (прогноз, описание или описание и прогноз); - последовательную адаптацию к нарушениям путем применения соответствующих вычислительных процедур; - повторные проверки нарушений и ранжирование при необходимости.

Основными трудностями при практической реализации РМ–подхода являются: - выбор глобального (или интегрированного) критерия оптимальности; - удовлетворительное решение задачи структурной идентификации в условиях большой размерности; - выбор оптимального маршрута проверки условий применения схемы РА – МНК и соответствующей адаптации. Для решения данных проблем и был разработан пакет СПОР [3, 8].

Основное назначение пакета – получение регрессионных моделей процессов, явлений или функционирования объектов с последующим их использованием для прогноза выходных характеристик (откликов) [9, 14]. Необходимость наличия подобной системы порождается большими затруднениями при выполнении подобных работ, требующих как многовариантности расчетов, так и применения различных методов оценивания параметров и структурной идентификации, а также анализа остатков при выбранном сценарии проверки соблюдения предположений МНК.

2. Адаптивное РМ

При разработке и использовании моделей прогноза основной целью является достижение свойств наилучшей линейной оценки (состоятельности, несмещенности, эффективности) для оценки прогнозируемой величины $\hat{Y}$. Эти свойства в первую очередь обеспечиваются подбором соответствующей (оптимальной по заданному критерию) структуры модели из множества (на основе постулируемой модели) конкурирующих структур. Таким образом, решается задача не только параметрической, но и структурной идентификации.

В подавляющем числе случаев постулируемая модель

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon; \quad i = \overline{i,n} \qquad (1)$$

не является оптимальной (адекватной) наблюдениям. Если считать линейную зависимость (1) подходящей, то основной проблемой будет размерность модели.

С одной стороны, опасаясь потерять существенные факторы, исследователь старается включить в правую часть модели (1) их как можно больше. Поэтому, как правило, модель является переопределенной, что приводит: а) к экономическим издержкам, б) к включению неинформативных, малоинформативных и дублирующих переменных.

Последнее приводит к возрастанию дисперсии прогноза $\hat{Y}$ для модели прогноза и к понижению точности оценивания β-коэффициентов в параметрической модели.

С другой стороны, недоопределенная модель, не содержащая значимых факторов, приводит к систематической ошибке Δ в прогнозе. И здесь возникает проблема соизмерения смещения Δ с величиной случайной ошибки прогноза в переопределенной модели. Чаще всего случайная ошибка оказывается большей, чем систематическая. К тому же в некоторых методах оценивания, отличных от МНК, модель намеренно обременяется смещением для уменьшения ошибки прогноза.

Таким образом, выдвигая гипотезу (1), исследователь сталкивается с множеством конкурирующих моделей (структур), содержащих $x_0 (x_0 = 1)$ и некоторое количество регрессоров из множества $\{x_1, \ldots, x_{p-1}\}$. Так как каждая переменная $x_j (j = 1, p-1)$ может либо входить в уравнение, либо нет, то всего получается 2^{p-1} моделей. Из этого множества структур необходимо выбрать по заданному критерию качества одну или несколько конкурирующим моделей.

Если используется стандартный регрессионный анализ (РА), то в прикладной статистике после анализа модели в целом и отдельных ее слагаемых прибегают к однокритериальному поиску оптимальной структуры. При невозможности применить полный перебор структур используют тот или иной известный вид неполного перебора по одному из критериев качества модели (средней квадратической ошибке σ, выборочному коэффициенту множественной корреляции *R*, *F*-критерию и т. д.).

Наиболее предпочтительными для структурной идентификации является ошибка на контрольной выборке. Данный критерий в максимальной степени отражает реальные случайные и систематические ошибки прогноза (отклика) и не имеет систематического хода по отношению к размерности. Ошибка на контрольной выборке, естественно, в той степени «истинна», в какой «истинны» контрольные значения y_i, обремененные, в свою очередь, разнообразными ошибками.

Применение подхода РМ требует разработки многокритериальных алгоритмов поиска. В общем случае для получения адекватной модели обработки данных необходимо решить многокритериальную задачу оптимизации путем последовательной адаптации к нарушениям условий РА–МНК.

В ряде случаев достаточно эффективными оказываются двухкритериальные методы. В рамках СПОР реализован метод пошаговой регрессии (схема включения с исключением), предусматривающий использование, помимо *F*-критерия, ряд других мер сравнения [10].

3. Критерии качества модели

Одной из важнейших задач при анализе данных является задача выбора критерия сравнения конкурирующих описаний.

В рамках СПОР предоставляется, помимо критериев качества модели на обучающей выборке (t, σ, F, R), возможность вычисления ошибки на контрольной выборке и ошибки на «скользящей» контрольной выборке [1, 19].

Качество модели РА обычно определяют по следующим критериям:

- средней квадратической ошибке σ, которая применяется как для оценки адекватности модели, так и для сравнения различных моделей друг с другом;

- выборочному коэффициенту множественной корреляции *R*, который используют как меру линейной связи (1): чем больше значение *R* ($0 \le R \le 1$), тем сильнее связь, то есть тем лучше аппроксимирующая функция соответствует наблюдениям, также высокое значение *R* гарантирует пригодность модели для прогноза;

- F-критерию, при $F > 4F_T(\alpha; p-1, n-p)$ (F_T – критическое значение, взятое из таблицы для *F*-критерия) модель признается заслуживающей внимания на предмет ее использования для прогноза.

Данные критерии качества характеризуют адекватность модели только по отношению к использованной для ее построения выборки точек (обучающая выборка). Это первый этап исследования модели, на котором экспериментатор должен быть убежден, что модель соответствует наблюдениям.

Если модель предназначена для прогноза, то надо быть уверенным в ее пригодности для определения области, не совпадающей с выборочными точками y_i.

Для оценки внешней адекватности (точности прогноза) используются контрольные точки. Исходная выборка делится на обучающую и контрольную. На первой выборке строится модель или множество моделей; на второй – выполняется оценка ее адекватности или дискриминация по статистикам.

Ошибка на контрольной выборке основана на анализе расхождений между прогнозом $\hat{Y}$ и известным наблюденным значением Y для объектов, не участвовавших в получении модели.

Поскольку, работая с малыми выборками, нет возможности разделить ее на обучающую и контрольную с достаточно большим количеством точек, для оценки внешней адекватности мы предлагаем использовать критерий, основанный на «скользящей» контрольной выборке. Если последовательно каждый из объектов выборки выводить из нее, полагая этот объект контрольным, и пересчитывать заново параметры модели, то разности между y_i и $\hat{y}_i$ для скользящей контрольной точки Δ_i = «наблюдение минус прогноз» ($i = \overline{1, n}$; где n – общее количество объектов) могут быть использованы для вычисления ошибки на «скользящей» контрольной выборке.

Последовательное исключение объектов, соответствующее удалению определенных строк из матрицы данных X, дает возможность сформулировать искусственно новую выборку (проверочную или контрольную) того же объема, что и исходная.

Все процедуры структурно-параметрической идентификации, включенные в пакет, реализуют вычисление рассмотренных статистик и поиск по ним оптимальной структуры модели. Более подробно критерии сравнения конкурирующих моделей рассмотрены в [19].

4. Программный пакет СПОР

При организации оптимальной предметной стратегии РМ необходимо учесть наличие различных выборок, предположений, классов функций, методов оценивания, мер качества и их наборов для принципа многокритериальности, методов структурной идентификации, конкурирующих в соответствии с принципом неокончательных решений стратегий адаптации к нарушению предположений.

Для практического применения РМ прежде всего требуется полная автоматизация всех заявленных процедур, для чего и было разработано соответствующее программное обеспечение.

Пакет СПОР включает в себя следующие модули:

1) управляющий модуль;
2) модуль формирования запроса;
3) библиотеку функциональных процедур;
4) блок сценария;
5) блок настройки системы;
6) блок редактора данных;
7) блок формирования таблиц;
8) справочник.

Основным инструментом положительного воздействия на прогностические свойства модели является алгоритм поиска ее оптимальной структуры. В пакете реализованы следующие процедуры структурно-параметрической идентификации:

- множественная линейная регрессия,
- гребневая регрессия,
- робастное оценивание,
- полный перебор структур,
- неполный перебор структур (перебор с ограничением на количество включаемых регрессоров в модель),
- перебор нормальных систем,
- пошаговая регрессия с включением-исключением,
- случайный поиск с адаптацией,
- случайный поиск с возвратом.

Данные процедуры могут быть вызваны как в автоматическом режиме для обработки целого ряда выборок данных, так и для обработки отдельной выборки данных по реализованному оптимальному сценарию [18].

В пакете реализована процедура построения и анализа графика остатков, что является полезным статистическим инструментом для проверки адекватности оцененной модели регрессии имеющимся данным.

Конкурентоспособность СПОР с другими статистическими пакетами:

• использование новых методов структурной идентификации: полный перебор, неполный перебор переопределенных и нормальных систем, многокритериальный метод пошаговой регрессии с включением-исключением;

• использование гибкого инструмента построения сравнительных таблиц;

• использование, помимо классических критериев качества модели на обучающей выборке, критериев качества модели на контрольной выборке и ошибки на «скользящей» контрольной выборке, что позволяет осуществлять оценку внешней адекватности модели (точности прогноза).

В настоящее время ведутся работы по наращиванию функциональных возможностей СПОР и ее интеллектуализации [15, 17].

5. Использование пакета СПОР

Пакет СПОР может использоваться для решения задач метода наименьших квадратов (задач восстановления зависимостей по избыточным косвенным наблюдениям) и регрессионного анализа в любых областях (экология, технологические процессы, экономика, социология и т.д.), различных задач, требующих восстановления эмпирической зависимости между выходным параметром процесса и набором входных.

При обработке аэрокосмических снимков [2] и решении ряда фотограмметрических задач [4] применение СПОР путем вычислительных экспериментов позволило получить следующие результаты:

1. Получение моделей преобразования координат по малым выборкам с дисперсией оценки точности в 1,2 – 100 раз меньшей дисперсии при стандартном подходе, что соответствует повышению точности аппроксимации при применении РМ от нескольких десятков процентов до одного порядка.

2. Повышение точности при использовании РМ обеспечивается процедурой структурной идентификации; реализация последней предполагает формирование множества конкурирующих структур на основе исходной перспективной модели и поиск оптимальной структуры по заданному критерию качества.

3. Поиск модели, оптимальной по ошибке на «скользящей» контрольной выборке приводит к модели с лучшими прогностическими свойствами по сравнению с моделями, оптимальными по средней квадратической ошибке и позволяет решить задачу выбора информативного по *t*-критерию набора регрессоров. В 70% всех случаев модели, оптимальные по средней квадратической ошибке, содержат малоинформативные слагаемые. Модели, оптимальные по ошибке на «скользящей» контрольной выборке, лишь в 17% всех случаев содержат незначимые слагаемые. Модели, содержащие малоинформативные слагаемые, полученные по ошибке на «скользящей» контрольной выборке, содержат по одному незначимому регрессору, в то время как модели, полученные по средней квадратической ошибке, как правило, два и более малоинформативных слагаемых. Было оценено, значимое ли улучшение по внешней точности дает использование в качестве критерия качества ошибки на «скользящей» контрольной выборке. Анализ показал, что использование данного критерия дает значимое улучшение прогностических свойств по сравнению со средней квадратической ошибкой. Устойчивость выводов по отношению к наблюдениям, вошедшим в контрольную выборку, была проверена и подтвердилась по 10 случайным экспериментам для каждого из трех произвольно выбранных снимков.

Пакет СПОР успешно применялся для обработки лазерных [5] и радиоинтерферометрических данных большой размерности [6, 7], для оценки качества питьевой воды [11, 16], для обработки социально-экономических показателей [12, 13].

6. Заключение

Пакет СПОР может оказаться полезным при разработке моделей прогноза как в высокоточных областях знаний и технологических процессах с входными характеристиками, содержащими взаимозависимые, неинформационные или малоинфомационные факторы, так и при описании социально-экономических явлений и экологических ситуаций. Применение пакета обеспечивает повышение точности прогнозирования при использовании оптимальной модели от нескольких десятков процентов до одного порядка.

Литература

[1] Валеев, С.Г. Система поиска оптимальных регрессий: учебное пособие / С.Г. Валеев, Г.Р. Кадырова. – Казань : ФЭН, 2003. – 160 с.

[2] Кадырова, Г.Р. Регрессионные модели для трансформации изображений на аэрокосмических снимках / Г.Р. Кадырова, Н.А. Билибина, Л.М. Бугаевский, С.Г. Валеев // Известия Вузов. Сер.: Геодезия и аэрофотосъемка. – 1997. – № 1. – С. 56–66.

[3] Валеев, С.Г. Автоматизированная система для решения задач метода наименьших квадратов / С.Г. Валеев, Г.Р. Кадырова // Известия Вузов. Сер.: Геодезия и аэрофотосъемка. – 1999. – № 6. – С. 124–130.

[4] Валеев, С.Г. Оптимальные редукционные модели в фотографической астрометрии / С.Г. Валеев, Г.Р. Кадырова // Известия Вузов. Сер.: Геодезия и аэрофотосъемка. – 2002. – № 3. – С. 58–69.

[5] Валеев, С.Г. Метод ступенчатой ортогонализации базиса и его применение при решении задач МНК / С.Г. Валеев, Т.Е. Родионова // Изв. вузов. Геодезия и аэрофотосъемка. – 2003. – № 6. – С.3–14.

[6] Валеев, С.Г. Методика статистической обработки РСДБ-наблюдений / С.Г. Валеев, Т.Е. Родионова, В.Е. Жаров // Изв. вузов. Геодезия и аэрофотосъемка. – 2008. – № 1. – С. 13–18.

[7] Валеев, С.Г. Вычислительные эксперименты по обработке РСДБ- наблюдений / С.Г. Валеев, Т.Е. Родионова, В.Е. Жаров // Изв. вузов. Геодезия и аэрофотосъемка. – 2008. – № 2. – С.94–100.

[8] Валеев, С.Г. Программная система поиска оптимальных регрессий / С.Г. Валеев, Г.Р. Кадырова, А.А. Турченко // Вопросы современной науки и практики. Сер. Технические науки. – 2008. – № 4(14), т. 2. – С. 97–101.

[9] Кадырова, Г.Р. Оценка и прогнозирование состояния технического объекта по регрессионным моделям регрессий // Автоматизация процессов управления. – 2015. – № 4(42). – С. 90–95.

[10] Кадырова, Г.Р. Модификация метода пошаговой регрессии для получения математических моделей прогноза поведения объекта // Автоматизация процессов управления. – 2016. – № 3(45). – С. 65–70.

[11] Родионова, Т.Е. Применение адаптивного регрессионного моделирования для описания функционирования технического объекта // Известия Самарского научного центра Российской академии наук. – 2014. – Т. 16. – № 6-2. – С. 572-575.

[12] Родионова, Т.Е. Применение математического моделирования для анализа влияния социальной сферы на качество жизни населения (на примере Ульяновской области) / Т.Е. Родионова, М.В. Рыбкина // Экономический анализ: теория и практика.– 2014. – Вып. 32(383). – С. 61-66.

[13] Родионова, Т.Е. Исследование влияния инфляции на социально-экономические факторы / Т.Е. Родионова, М.В. Рыбкина, Н.А. Ананьева // Качество. Инновации. Образование. – 2015. – №9(124).– С. 48-51.

[14] Кадырова, Г.Р. Программная система поиска оптимальных регрессионных моделей прогноза // Путь науки. – 2014.– № 7 (7). – С. 10–11.

[15] Кадырова, Г.Р. Система поиска оптимальной модели. Состояние дел и перспективы развития // Потенциал современной науки. – 2015. – № 4 (12). – С. 8–10.

[16] Родионова, Т.Е. Статистические методы оценки показателей качества питьевой воды / Т.Е. Родионова, В.Н. Клячкин // Доклады АН ВШ РФ. – 2014. – №2-3 (23-24). –С. 101-110.

[17] Кадырова, Г.Р. Возможности программной системы регрессионного моделирования для оценивания модели и поиска ее оптимальной структуры // Радиоэлектронная техника. – 2015. – № 2 (8). – С. 228–233.

[18] Кадырова, Г.Р. Формирование стратегий поиска оптимальных регрессий // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. – 2016. – № 1 (10). – С. 178–180.

[19] Кадырова, Г.Р. Исследование мер качества моделей для оценивания состояния технического объекта // Синтез, анализ и диагностика электронных цепей. – 2016. – Вып. 13. – С. 71-83.